Marking consistency metrics

November 2016

Ofqual/16/6121

Authors

This report was written by Stephen Rhead and Beth Black from Ofqual's Strategy Risk and Research directorate, and Anne Pinot de Moira, consultant.

Contents

1 Executive summary

In 2014, Ofqual committed to develop a set of metrics to measure the quality of marking in general qualifications. It was envisaged that such metrics should help us to better monitor and quantify the quality of marking in general qualifications.

The purpose of this report is to present some technical work describing some potential metrics. Accordingly, while all the metrics presented are based upon real data gathered from exam boards, it is not possible to identify any particular unit from any particular exam board. In due course, this data will help us to develop how acceptable levels of marking consistency can be established for different assessment types.

The following technical report gathers information from four exam boards (AQA, OCR, Pearson and WJEC) and presents some technical work describing some potential metrics.

This report is in 3 sections. The first section describes the sources of the data used for the marking metrics, namely data generated as a product of the onscreen marking monitoring processes employed by the exam boards. This section describes areas of similarities and differences between the processes and therefore the data available for generating the metrics. A number of assumptions required for the various onscreen monitoring data and the derivation of metrics are outlined.

The second section presents a number of possible marking consistency metrics at different levels of granularity: question (item) level metrics; component/unit level metrics and potential qualification level metrics. Due to the prevalence of segmented marking where candidates' scripts are distributed to multiple markers for item level marking, the metrics at component and qualification level are necessarily derived and built up from item level marking consistency data.

In the third section, a series of caveats are presented; most notably it is essential that the use of suitable marking metrics does not compromise the live online monitoring process. Lastly some areas for further work are suggested.

In summary, this report represents the first stage of this work in deriving marking metrics. There are important areas around the practical usage of these metrics which need careful consideration. These areas include: how acceptable levels of marking consistency can be established for different assessment types; and how such metrics can be used to drive improvements in marking.

2 Introduction

In 2014 Ofqual published a report on the quality of marking for A levels, GCSEs and other qualifications (Ofqual, 2014). This report presented an in-depth review of the current marking system and set out a series of recommendations to improve the quality of marking of examinations. One recommendation was to develop a set of metrics to monitor the quality of marking of general qualification types.

This report sets out a number of proposals for the derivation of quality of marking metrics. In order to derive the metrics, a brief overview of the monitoring procedures used by the different exam boards will be given. A series of item level statistics will be derived and used as the building blocks for component level metrics. These metrics are scaled to specification level to illustrate how they could potentially be used for linear qualifications as reformed GCSEs and A levels are phased in.

Finally, this report sets out the limitations and necessary data assumptions, as well as highlighting the potential impact of metrics on the live marker monitoring process.

In summary, this report represents the first stage of this work in deriving marking metrics. There are important areas around the practical usage of these metrics which need careful consideration in the near future. These areas include: how acceptable levels of marking consistency can be established for different assessment types; and whether and how such metrics can be used to drive improvements in marking.

3 Marker monitoring in onscreen marking and the data produced

3.1 Marker monitoring in onscreen marking

All four of the exam boards (AQA, OCR, Pearson and WJEC) who provided marking data for this project use onscreen marking, and they monitor marking quality during the marking session. This produces an electronic record of the monitoring of quality of marking. Onscreen marking is mainly monitored using one of two procedures. The first and most common approach is the introduction of pre-marked responses into an examiner's script allocation. These pre-marked responses are known as seed(ing) items or sometimes validity items. From here on we will refer to these as seed items. Seed items are introduced at times and intervals generally unknown to the examiner. Sampling rates of approximately 5% are typical. The examiner marks the item 'blind', i.e. unaware that it is a seed item and without sight of the pre-determined mark. A

comparison of the two marks derived from this process allow an assessment of the examiner's marking against a pre-agreed standard. This process is illustrated in figure 1.

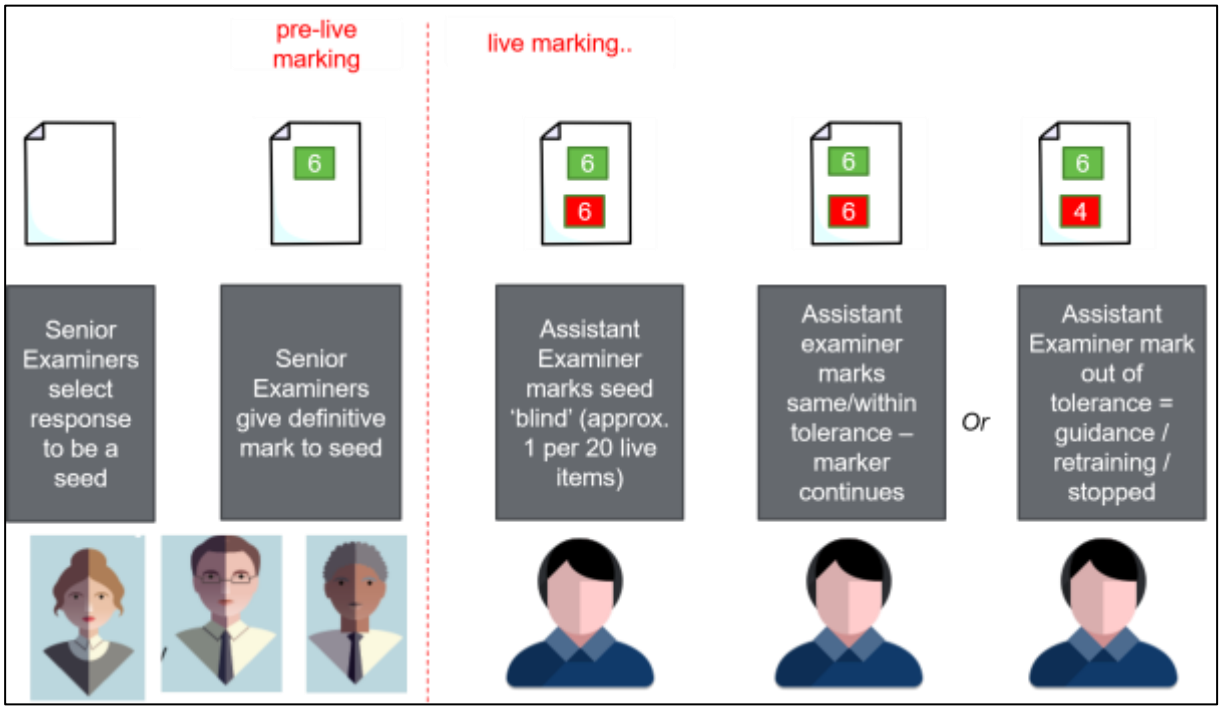

Figure 1. *The seeding process. Prior to live-marking, senior examiners select responses to be seed items and assign a definitive mark to the seed item. The definitive mark awarded to the seed is that which will contribute to the final mark of a candidate. These pre-marked responses are introduced into an assistant examiner's allocation at intervals and times unknown to the examiner. The mark awarded by the assistant examiner does not contribute to the candidate's final mark and is used as a mechanism to monitor marking. If the assistant examiner's mark agrees with the definitive mark or is within tolerance, the examiner can continue to mark, if the mark is out of tolerance the examiner may be given guidance, retraining or stopped from marking.*

All exam boards in this study have onscreen marking systems that allow monitoring by seeds. However, the exam boards have differing approaches to allocation of seed items. Some boards and marking systems distribute seeds at item level or groups of items, whereas some boards and marking systems distribute seeds only at the level of script rather than item. In this latter case, for any single examiner the seed is therefore the entire pre-marked script but item level information is still captured. In both systems (whole script or item seeding) the final mark for the seed item which contributes to the candidate's overall mark is known as the 'definitive' mark.

There are many ways for arriving at a single, definitive, mark of a seed item (see Tisi, Whitehouse, Maughan, & Burdett, 2013), although typically once the seeds have been selected, the definitive mark is generally derived by one or more senior examiners, often but not always including the Principal Examiner for that unit. Exam boards generally allow some flexibility and there is no formal record for each seeding item of precisely who was involved in recording the definitive mark. In order to incorporate seed items in the derivation of quality of marking metrics, it has been necessary to assume that the way in which the final mark is derived introduces no bias to potential quality of marking metrics and to accept the seed mark as the definitive mark no matter how it was derived.

Along with seeds, some boards also employ a system of blind sample-double marking which is typically used for an extended response (illustrated in figure 2). In this approach a series of randomly chosen responses will be blind marked by two randomly paired examiners. For all boards the examiners are chosen from the entire pool of examiners. However, how the final mark is awarded to the candidate varies by board. In one approach the final mark awarded to a blind sample double-marked response is the higher of the two marks unless they differ by more than a pre-agreed tolerance (the 'higher mark' approach). For the second approach the second examiner is always a senior examiner and the final mark awarded is that of the senior examiner (the hierarchal approach).

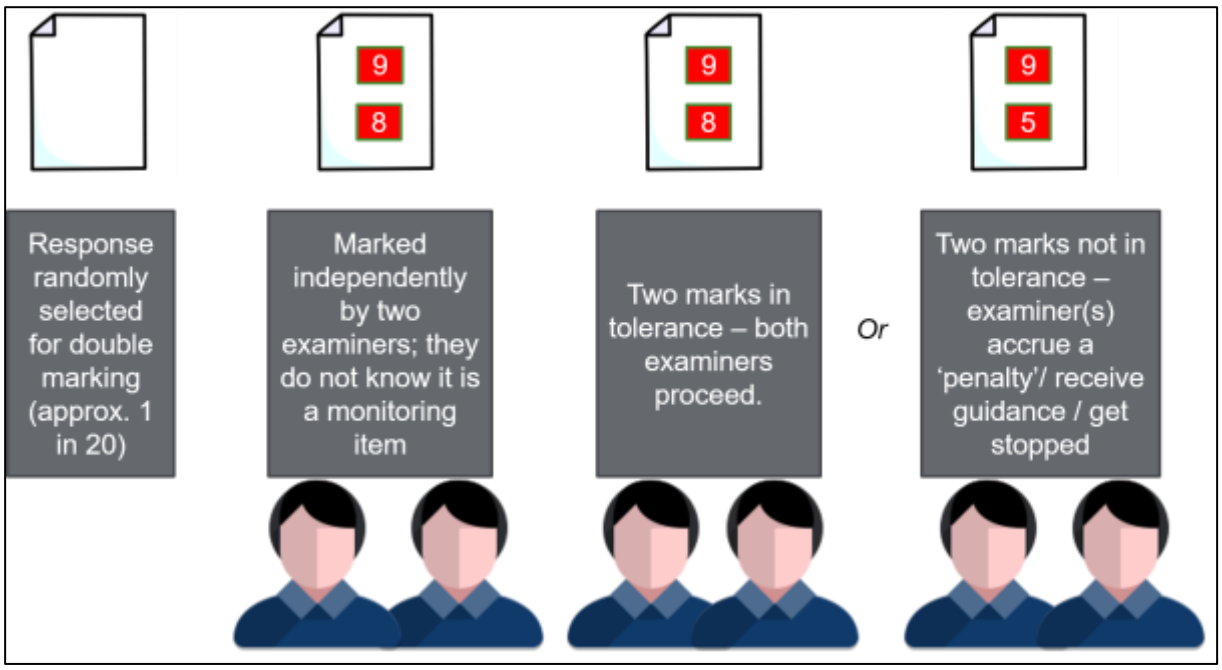

Figure 2. *The process behind blind sample-double marking.*

Regardless of either approach used (seed items or blind sample double-marking), the two marks awarded to a single response were arrived at independently of one another and as a result can be treated as independent in the statistical sense (Bramley & Dhawan, 2010).

3.2 Data produced from monitoring marking

In this report, marking metrics are created from the data arising from the operational monitoring of quality of marking during the live marking session. It has been assumed that the most appropriate measure of quality of marking is based on the difference between two marks given for a single response. Thus the data used in the project is the mark-remark data.

Mark-remark data for all items on all online marked units was requested for the following subjects: business studies, English language, English literature, French, geography, history, physical education, physics, psychology, sociology and Spanish from four exam boards, AQA, OCR, Pearson and WJEC. Data at GCSE and GCE level was requested. These were chosen in order to represent a range of subjects, item types and examination structures.

This data set has 433 unique units/components and some 66.7 million items; of which approximately 11.8 million were seed items, 600,000 sample double-marked items and 54.5 million automarked items (typically multiple choice, objective response or one-word response items which can be computer-read). There were no discrepancies between the initial mark and the final mark for any auto-marked item.

With the exclusion of automarked items, each item in the dataset has marks awarded by two or more examiners. This mark-remark data is the foundation of this analysis. For seed items the first examiner mark and the final mark awarded to the candidate are defined as the mark-remark data. Hierarchal sample-double marked items are analogous to this, the first examiner mark and the final mark awarded are defined as the mark-remark data. The final mark awarded to the item was missing for some of the ‘higher-mark approach’ sample-double marked data, consequentially the first examiner mark and second examiner mark were defined as the mark-remark data. The mark-remark difference is given by the following relationship:

$$difference = mark\ awarded\ by\ examiner\ 1 - final\ mark\ awarded\ .$$

A positive mark-remark difference means that the first examiner has awarded a mark more lenient than the definitive mark and negative difference corresponds to a more

severe mark. There are differences in the way that the seeds are chosen, the way that seed marks are derived and the way that examiner hierarchy is respected. It has been necessary to assume that the final mark awarded to an item is the definitive mark, regardless of how that mark was generated.

4 Metrics

In their 2010 report, Bramley and Dhawan present the idea of quality of marking as distinct from the reliability of assessment, describing the concept as examiner-related variability or examiner accuracy. With this is mind, the metrics presented here are all derived from the mark-remark data arising from multiple responses to the seed and sample double-marked items.

Ideally quality of marking metrics should be presented at the least granular level possible allowing comparisons between similar specifications. However, not all subjects or specifications are 100% externally assessed examinations and so the metrics suggested here are also presented at item and component level. These also have a valuable role in understanding marking consistency at this lower level of granularity.

Where on-screen marking is distributed at script level, derivation of component level metrics is relatively straightforward. However, as the majority of on-screen marking is segmented (i.e. distributed for marking at item level rather than script level), derivation of component level metrics is non-trivial. In such instances, component level metrics are derived from item level statistics for each question within a component (figure 3).

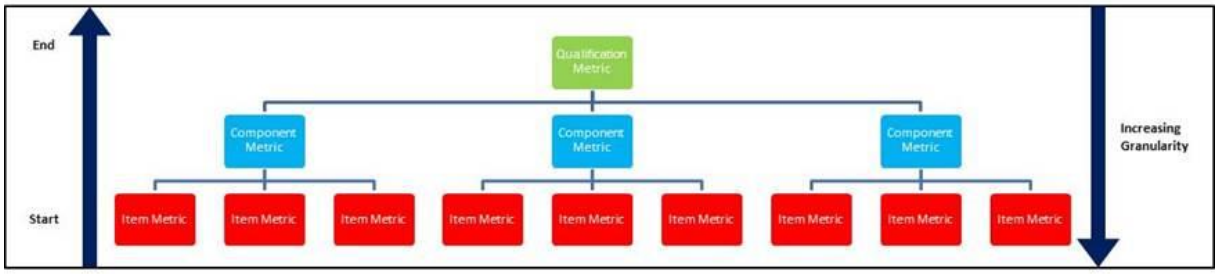

Figure 3. *The process of deriving component and qualification level metrics within a single qualification. Component level metrics are derived by the aggregation of item level statistics for all questions within a particular component. Likewise, qualification metrics are derived by aggregating over all components in a qualification. These statistics are complementary; a metric from one level may be used to contextualise information in another.*

Lastly, metrics need to be understood by the target audience. Whilst typically they may be presented as some form of mark difference, or probability of receiving the definitive mark, it may be desirable to contextualise quality of marking in terms of the position of the grade boundaries. This can help contextualise how quality of marking may affect a candidate's overall final grade – at component or at qualification level.

4.1 Item level statistics

For each question the mean and standard deviation of the mark difference can be calculated using the awarded mark; this may be presented in terms of raw marks (figure 4) or as a percentage of the maximum mark of the item (figure 5). These distributions are across all units for all subjects for both GCSE and GCE. It is observed that the standard deviation scales approximate proportionally with the maximum mark of the item.

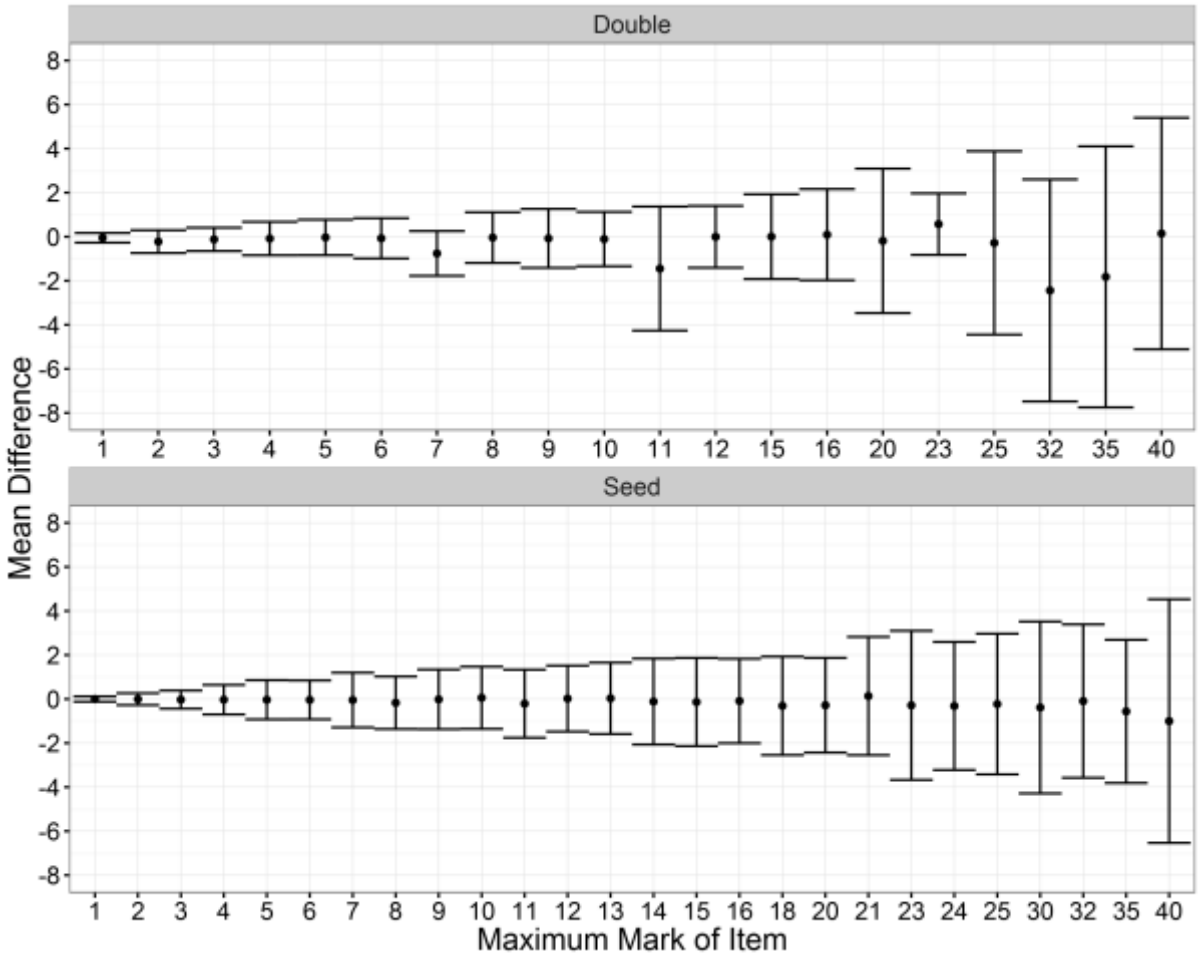

Figure 4. *Mean mark difference between the mark awarded by the first examiner and the final mark awarded to candidate. The mean mark difference is given by the solid black point and the standard deviation is given by the whiskers. The standard*

deviation is a measure used to quantify the amount of variation of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean value of the set, whereas a high standard deviation indicates that the data points are spread over a wider range of values.

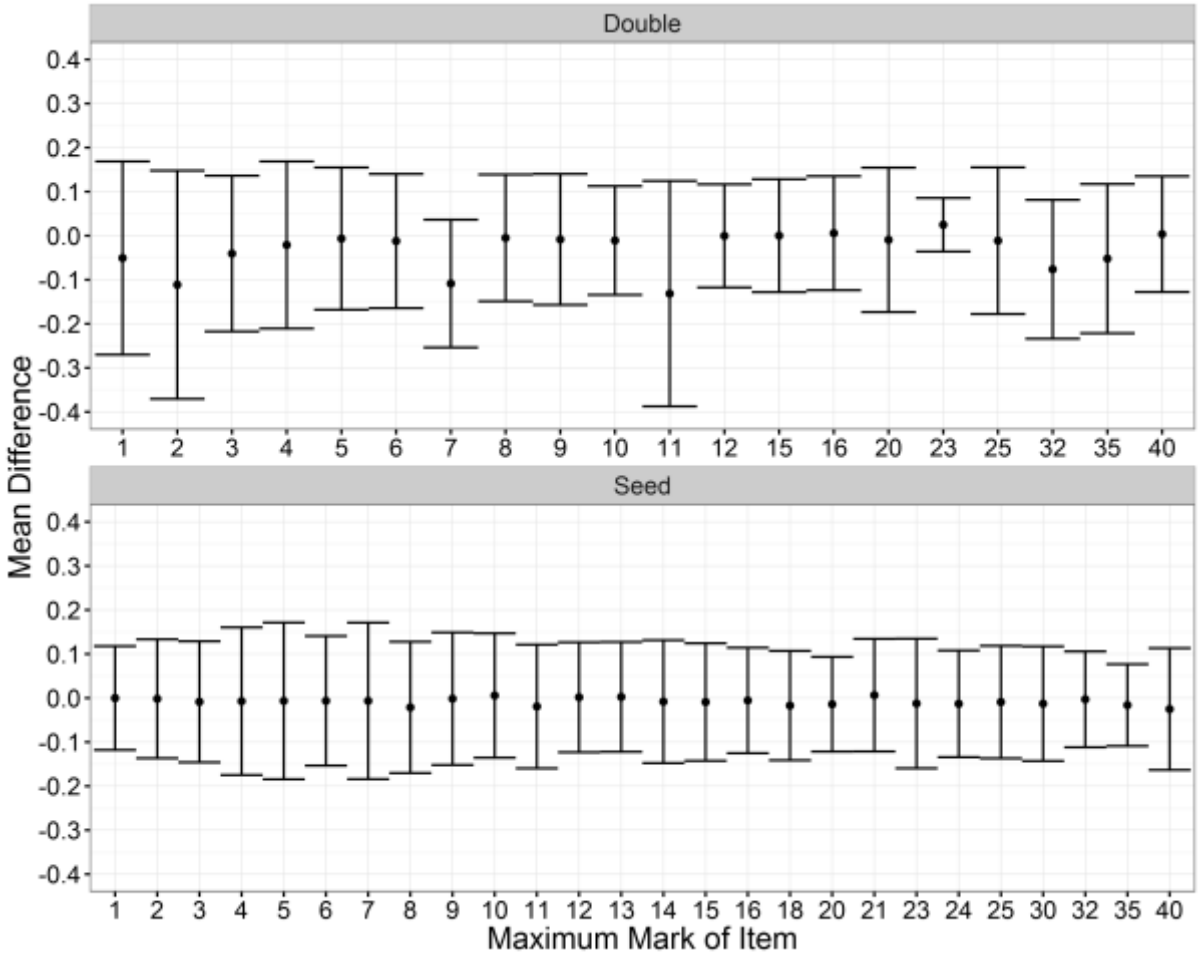

Figure 5. *Mean mark difference (scaled by the maximum mark of the item) between the mark awarded by the first examiner and the final mark awarded to candidate. It is observed that when scaled by the maximum mark of the item the standard deviations are approximately constant.*

It is also straightforward to calculate the probability of the exact agreement between the mark awarded by the first examiner and the final mark awarded (figure 6). The median probability (denoted by the black bar) generally reduces exponentially with the maximum mark of the item, with this effect particularly clear for seed items. This trend is not particularly surprising as the complexity and subjectivity of a question most likely increases with the maximum mark, leading to an increase in the likelihood of differences between examiners. It is important to stress that seeds with a high probability of agreement may or may not be ‘good’ seed items in terms of their

primary function; for example, it is possible that items with 100% agreement and high mark tariff are all null responses and as a result are marked perfectly accurately. Seeds with a low probability are not necessarily poorly marked or poorly chosen seeds. Many may well provide a good monitoring tool and expose problems with marking accuracy. If metrics are to be derived from on-screen marked data, it is important that all seeds which give information on marking quality performance remain in the pool so that the apparent quality of marking is not artificially inflated or deflated.

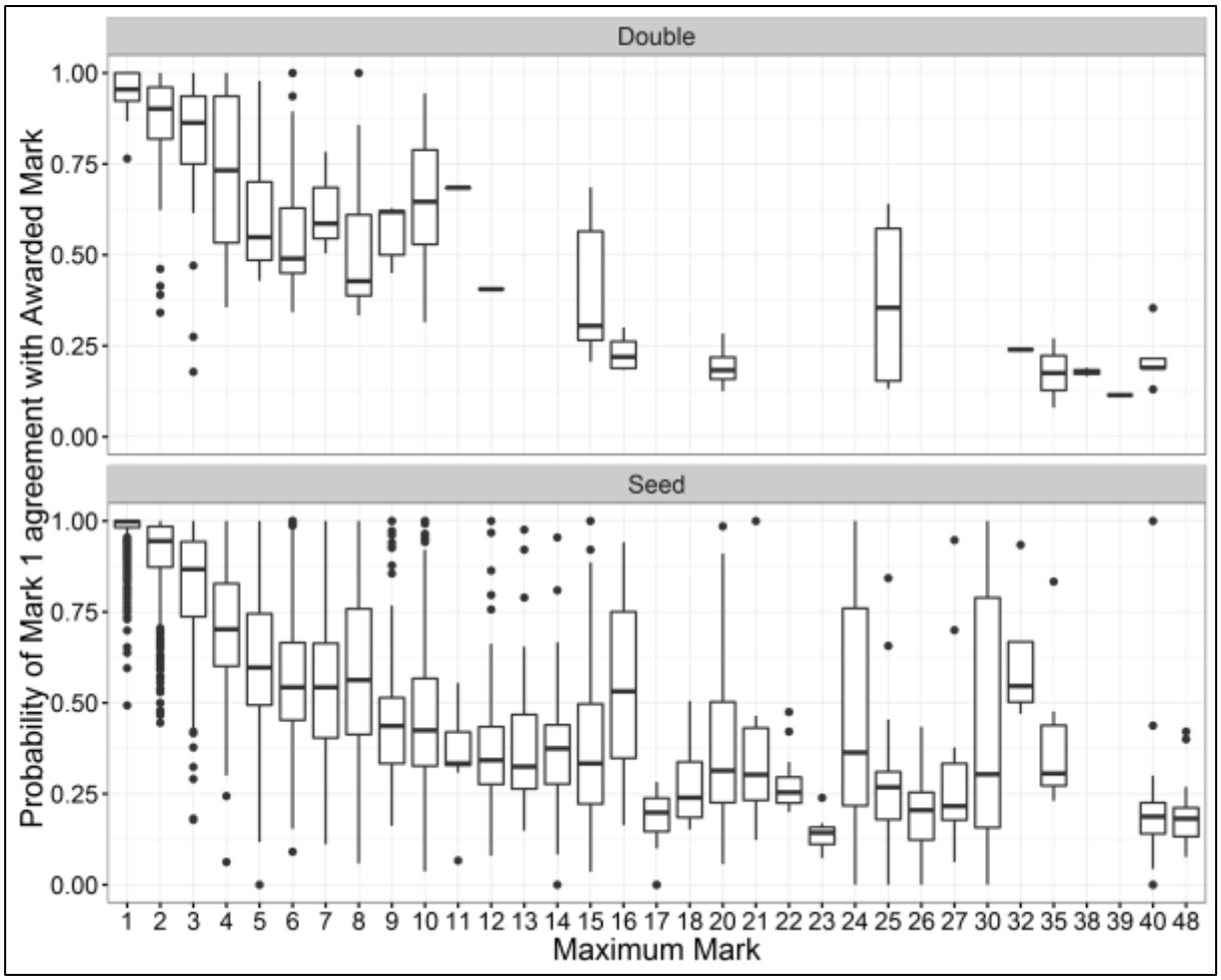

Figure 6. *Boxplot illustrating the agreement between the mark awarded by the first examiner and the final mark awarded. Boxplots are a standard way of displaying distributions of data. The median marks the mid-point of the data and is shown by the black line that divides the box into two parts; the box represents the interquartile range (the middle 50% of the data). The whiskers represent the data outside of the interquartile range and they extend 1.5 times the interquartile range from the top and bottom of the box respectively. The bigger the box and whiskers the greater the variability. Data that falls outside of the whiskers are known as outliers and are*

illustrated by the solid points. The outliers can provide a starting point for identifying potentially problematic items.

Distributions of mark differences from the final mark at item level could be used to identify any biases in marking; the distribution of mark difference for physics seed items is shown in figure 7 and tabulated in table 1 (this contains data from all exam boards and both GCSE and GCE). Typically, each physics question is relatively objective, has a low-mark tariff (generally ≤ 6 marks), and as a result the mark difference is found to be zero for nearly all seed items. Due to the almost perfect symmetry of the distribution around zero difference, on the whole the marking is found to show no bias towards severe or lenient marking.

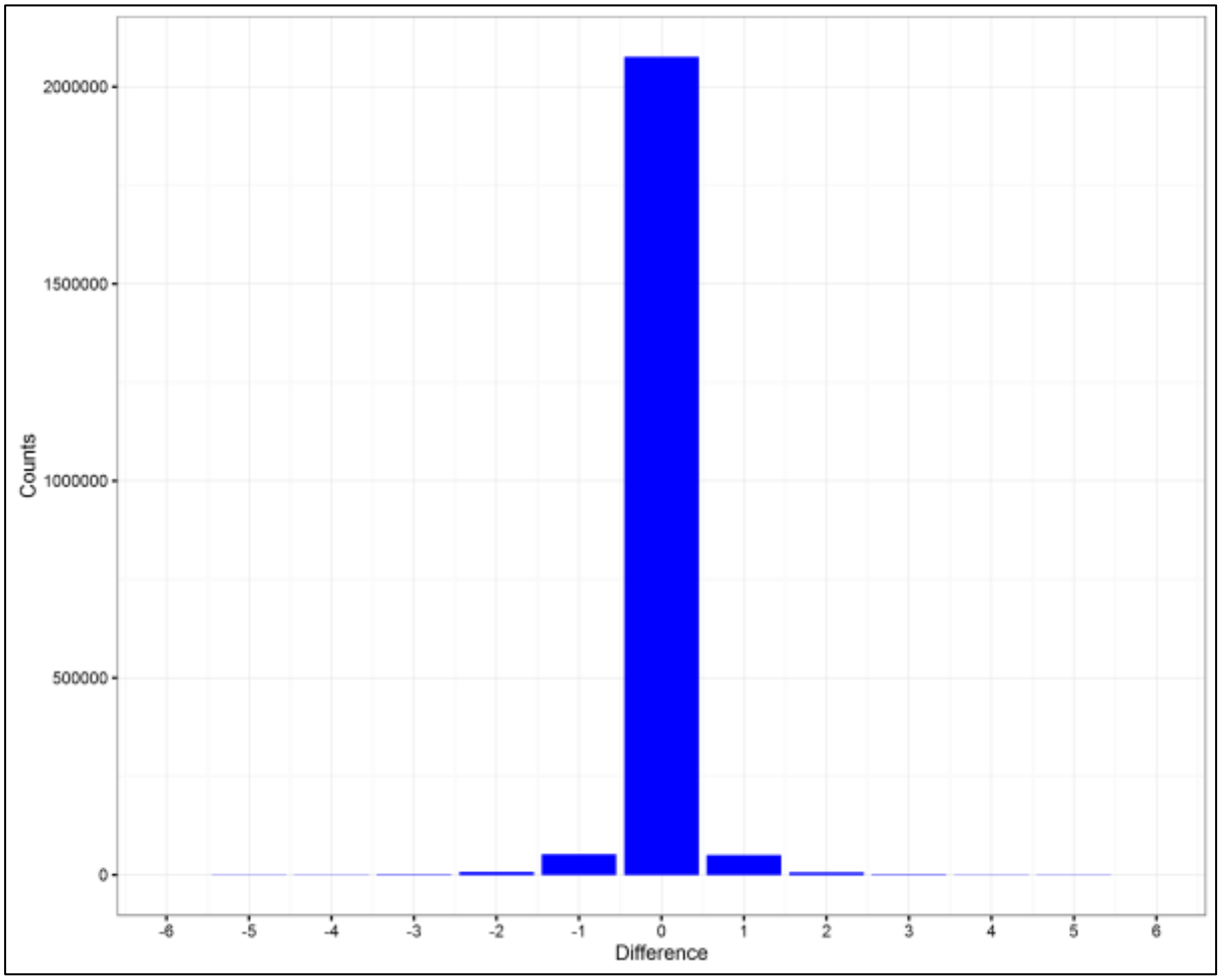

Figure 7. Distribution of mark difference from the awarded mark for all 2015 physics items.

Table 1. *Distribution of mark differences for all physics seed items.*

Mark Difference	2013	2014	2015
-6	1	4	7
-5	5	19	9
-4	69	200	90
-3	417	744	504
-2	4,364	6,983	6,646
-1	33,993	52,696	52,427
0	1,261,165	1,754,047	2,075,847
1	28,530	48,109	50,338
2	2,942	5,037	5,756
3	298	591	533
4	38	95	126
5	3	21	5
6	0	1	0

There are a number of simple statistics that can be derived at item level for the provided on-screen marked data. These statistics only provide information about quality of marking at item level. However, this data can be used to contextualise information at component level; if a particular component gives cause for concern then these metrics may be used to identify problematic items within this component. Looking at problematic items may also be instructive in future designs of assessments.

4.2 Component level metrics

Our attention now turns towards the derivation of component level metrics and, as only some boards/marking systems use whole script seeds, component level metrics are obtained by aggregating up from item level. Due to the majority of on-screen marking being segmented there are some questions that have no mark-remark data. This happened for one of two scenarios: (i) the questions were automarked and not included in the mark-remark data, or (ii) questions were missing and not automarked. In order to build a metric of quality of marking for these components it is necessary to reintroduce the missing questions to the dataset. It is possible to introduce missing automarked questions by assuming they have been marked perfectly accurately. The second scenario is more difficult to correct for. However, it is possible to substitute each missing item with the mean difference across the entire component although this would most likely lead to an over- or under-estimate of quality of marking (particularly if the missing items were more simple or more complex than the present items). In order to create a complete picture of quality of marking for a component, it is necessary to have information for all questions within that component (including automarked questions). As a result, automarked questions, where missing, were reintroduced as having perfect accuracy and analysis in this

report focussed on components where responses to all questions were present.

Sum of independent random variables

For each question within a component it is possible to calculate the mean and standard deviation of the difference from the awarded mark. From this, an estimate of the mean and standard deviation at component level can be obtained. If $E(X_1)$, $E(X_2)$, … , $E(X_n)$ are the expected difference for each of the n questions within a component and $X_1, X_2, \ldots X_n$ are random variables with known distributions, the expected difference from the awarded mark at component level may be given by:

$$E(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} E(X_i) \; . \qquad (1)$$

Likewise, the variance at component level can be expressed by:

$$V(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} V(X_i) + \sum_{i \neq j} \mathrm{cov}\left(X_i, X_j\right) \approx \sum_{i=1}^{n} V(X_i) \; , \; (2)$$

where $V(X_n)$ is the variance of the n^{th} question within a component. Due to the segmented nature of the majority of on-screen marking, it has been assumed that the distribution of differences between questions are independent; as a result, the covariance term is zero[1]. The standard deviation is obtained by taking the square root of equation 2.

Use of equations 1 and 2 allows an estimate of quality of marking, in terms of the mean difference at component level, to be obtained. This is illustrated in figure 8 for all physics components (the components have been anonymised and randomised). The expected difference from the awarded mark at component level is found to be within ± 1.5 marks and the standard deviation within ± 4 marks for all components. On average, the expected difference is close to zero, suggesting that examiners for each component show no systematic bias towards severe or lenient marking.

A drawback of comparing the expected differences across various components is that components vary in a number of dimensions; in particular, in terms of the maximum mark, the number of items and the distribution of scores. This means that comparisons using raw marks are not necessarily “like for like”. If components were on a common scale then meaningful comparisons could be made; a possible solution

[1] However, if the same person has marked the entire seeding script or all of a subset of items this may not be the case were they systematically lenient or severe.

is to contextualise the expected difference in terms of the maximum mark of the component. By taking such an approach, marker accuracy is given as a percentage of the maximum mark and the reported statistic would be the expected difference and standard deviation expressed as a percentage.

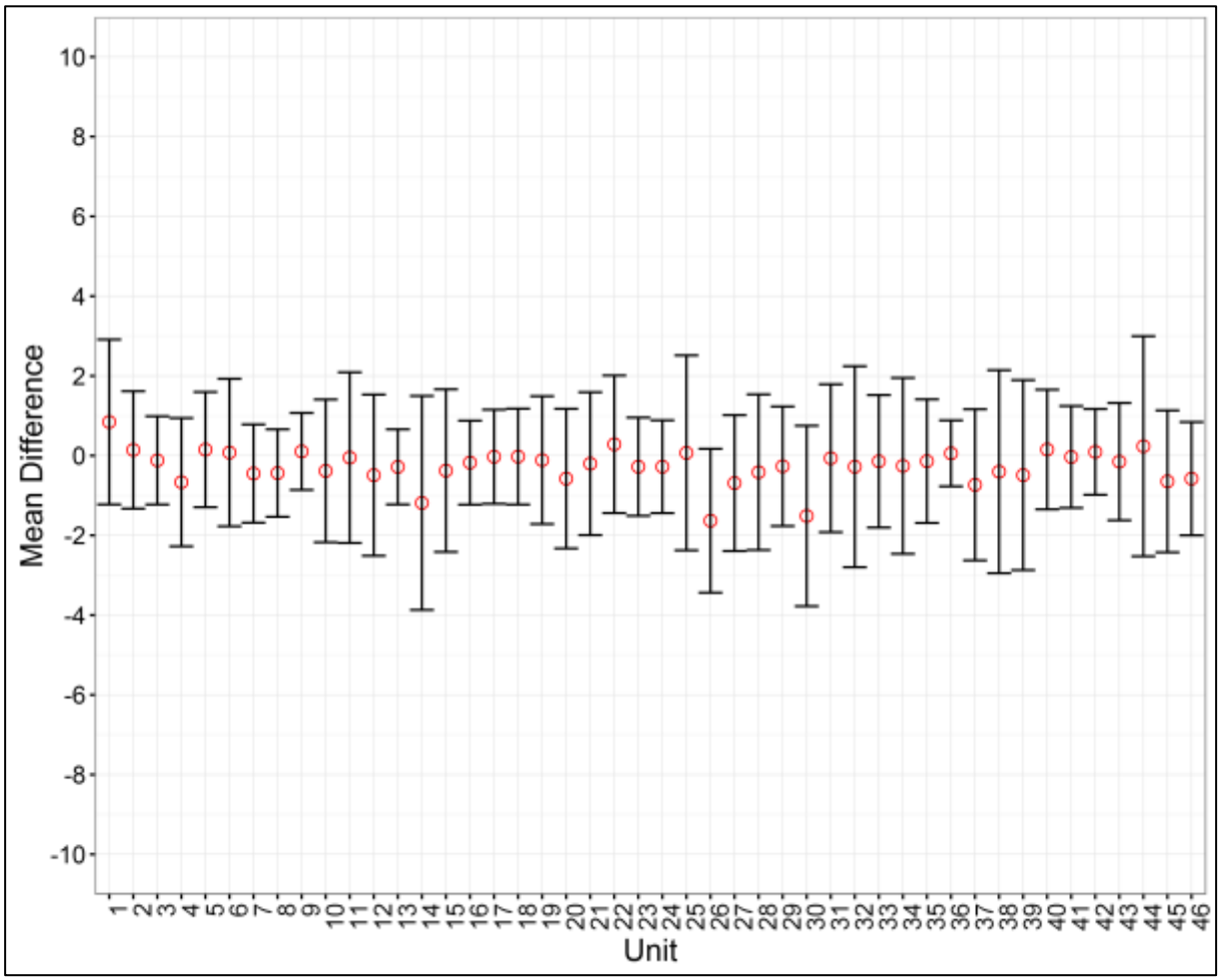

Figure 8. *The mean difference (red open circle) from the awarded mark expressed in raw marks for each physics component. The standard deviation is denoted by the black whiskers.*

Such an approach is illustrated in figure 9. The expected difference and standard deviations for all components are found to be within 2% and 5% of the maximum mark of the component respectively. By use of this metric we can see that marker agreement is generally similar for all physics components regardless of board or level. This metric has the advantage of being transparent and easy to understand in terms of its construction. Also, once the data is in this form it is easy to contextualise in terms of, for example, particular judgemental grade boundary widths (if desirable).

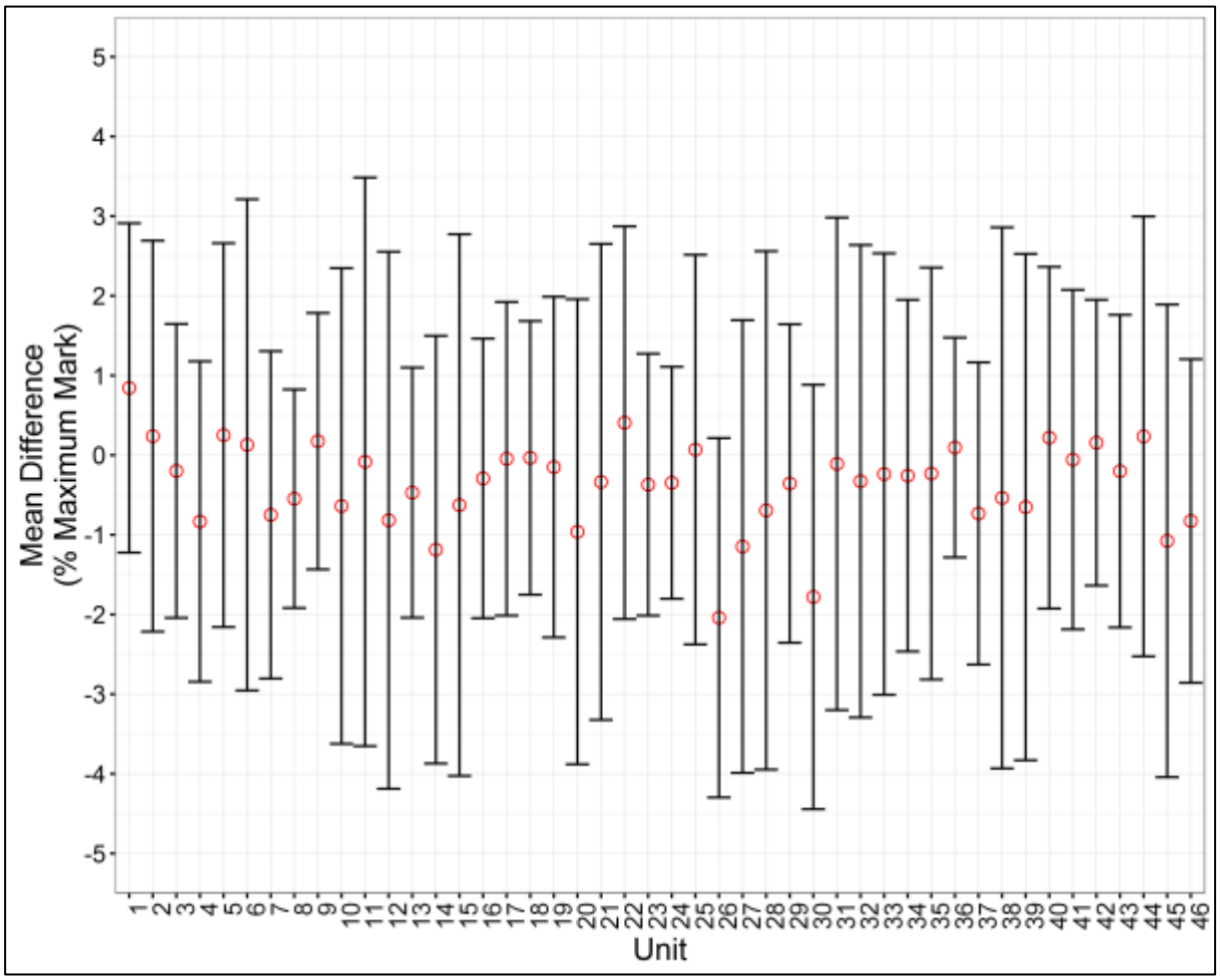

Figure 9. *The mean difference from the awarded mark expressed as a percentage of the maximum mark for each physics component*.

Pseudo-candidates

An alternative approach of presenting the expected difference could be obtained by predicting the difference from the awarded mark for a set of randomly generated candidates. These pseudo-candidates would have simulated whole script responses for all questions within a component, allowing the derivation of a component level metric, even in instances where on-screen marking is segmented.

Using equations 1 and 2, the expected difference and standard deviation are calculated for each component. The distributions of differences are calculated using these parameters to simulate a random normal distribution of differences for 150,000 candidates for each component. Each candidate has an estimate of the accuracy of marking expressed as a difference from the awarded mark. The number of candidates at each mark difference is converted into a percentage of the total number of candidates.

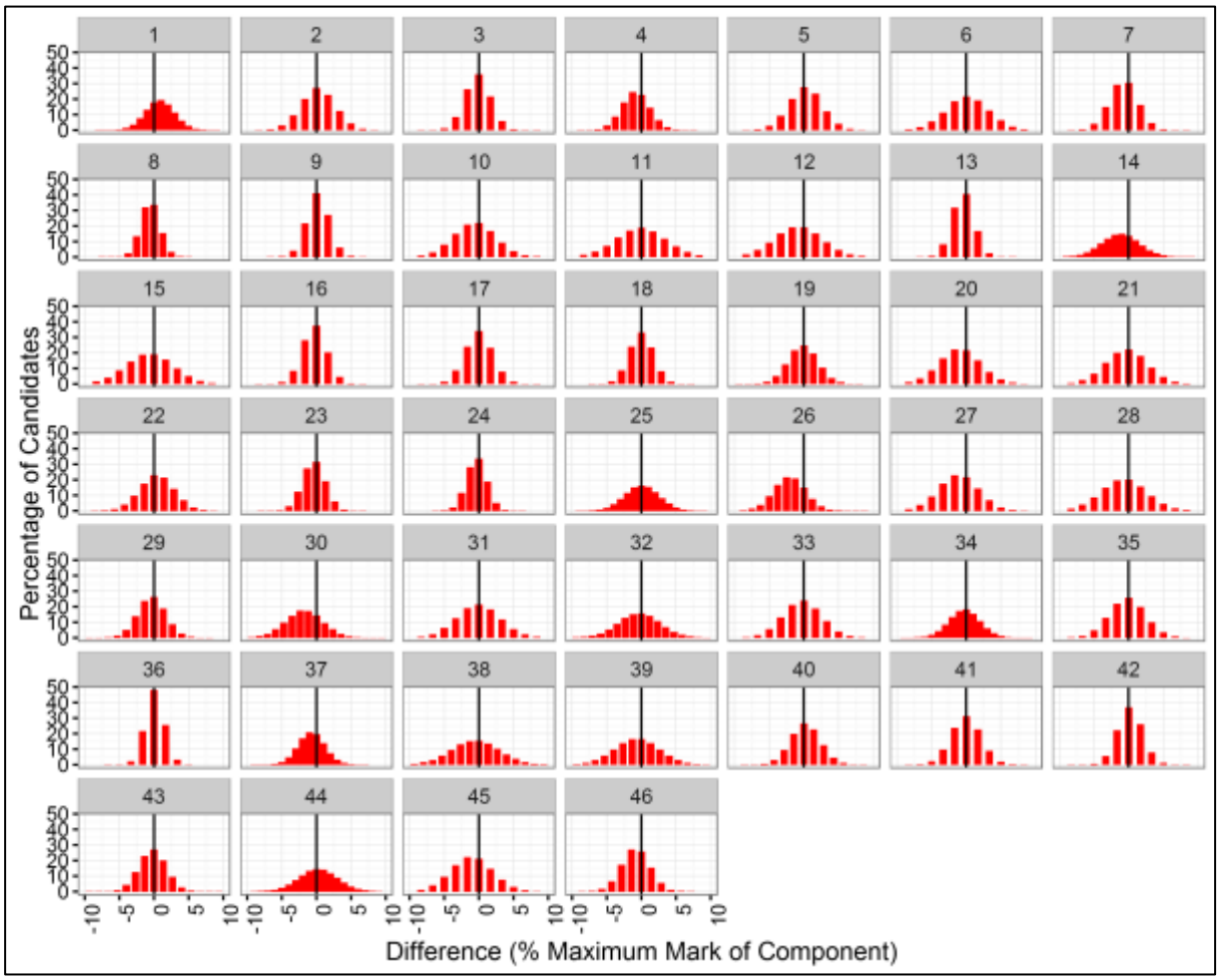

Figure 10. *The simulated difference from the awarded mark for each physics component based on the randomly generated candidates.*

The simulated difference for the randomly generated candidates are shown in figure 10. Again the differences are expressed as a percentage of the maximum mark of the component to mitigate for differences in the maximum mark and number of items for each physics component. These distributions may be used to demonstrate the effect that changing expected differences and standard deviations have on the distribution of differences at script level. The output statistic would be the expected difference and standard deviation of the distributions.

It might be instructive to tabulate these distributions, allowing for quick comparisons of the cumulative probability of a pseudo-candidate falling within a particular number of marks of the definitive mark; such distributions for units 36 and 6 are shown in table 2 for illustrative purposes (both components are worth 60 marks). This metric would allow for comparisons with marking tolerances. For example, table 2 shows how the percentage of candidates that fall within so many mark differences from the definitive mark. It could then be decided if these distributions are acceptable or not.

Table 2. *Distribution of the mark difference from the definitive mark for pseudo-candidates.*

Mark Difference (Raw)	Mark Difference (% Maximum mark)	Percentage of Candidates (Unit 36)	Percentage of Candidates (Unit 6)
0	0	48.0	21.5
± 1	± 1.7	94.6	58.7
± 2	± 3.3	99.9	82.8
± 3	± 5	100	94.4
± 4	± 6.7	100	98.6
± 5	± 8.3	100	99.7

The pseudo-candidates are derived using the expected difference for all items within a component. Key to trusting the simulated differences is to validate these results with actual differences obtained directly from the whole script marking data where available. A comparison of script level difference between pseudo and actual candidates is shown in figure 11; the high level of agreement between the two suggests pseudo-candidates may be used in the absence of whole script marking.

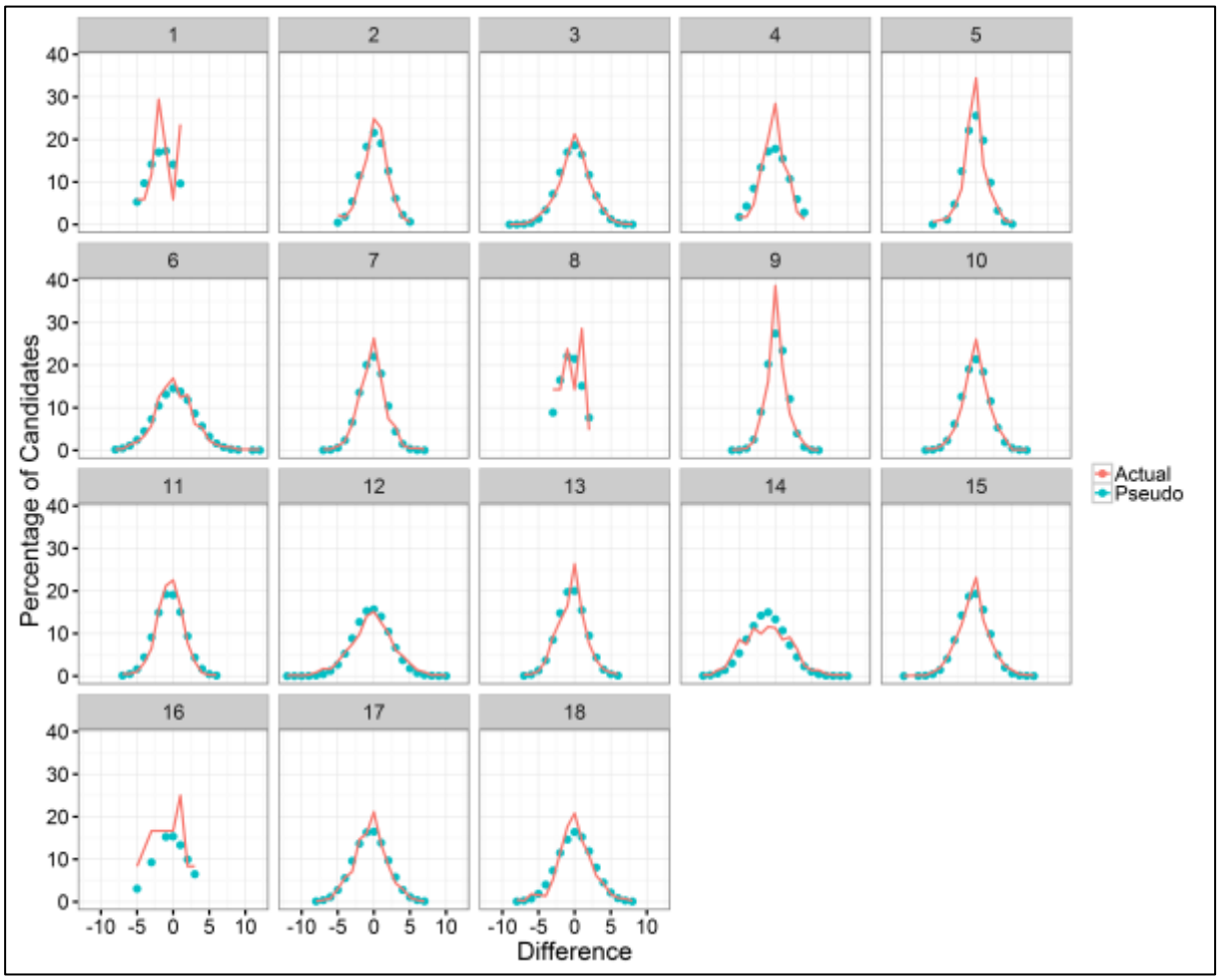

Figure 11. *Comparison of the simulated difference against the difference obtained directly from the whole script data. NB The simulated and actual data are in very good agreement with one another.*

Probability of definitive ('true') grade

As previously mentioned, components need to be on a common scale to allow for meaningful comparisons; the grade scale is one such common scale and as a result the creation of a metric which relates quality of marking to grading may be desirable. This may also be appropriate from a public understanding perspective, as the grade is the key information reported to the examinee. These metrics would contextualise quality of marking in terms of grading.

While there may be other potential conceptions of 'true' grade, in this section we are referring to the grade which would be derived from the definitive marks assigned to the seed items.

Using the expected difference and standard deviation, the probability of a particular mark resulting in the definitive grade classification can be calculated using the distance to the nearest grade boundaries as cut points on the normal distribution (figure 12). For each final mark awarded to a candidate, the black line represents the probability that the candidate has been awarded the definitive grade. The probabilities dip in the mark region near the grade boundaries and are highest at the extremes of the mark distribution. To a large extent the probability that a candidate is awarded the definitive grade is determined by their mark position relative to the grade boundary; a script with a 'true mark' exactly on the grade boundary but which is marked severely or leniently by a single mark is at greater risk of not receiving the 'true' grade than one with its 'true' marks several marks away from any grade boundary. The influence of quality of marking impacts in one of two ways: (i) the extent to which the probability dips at the grade boundaries and rises in between grade boundaries is determined by the standard deviation and expected difference and (ii) the expected difference affects the symmetry between grade boundaries. Negative expected differences lead to higher probabilities at the upper end of the grade boundary as it is less likely a candidate has been over-graded. The reverse is true for positive expected differences, where probabilities are higher at the lower end of the grade boundary given that it is less likely that a candidate has been under-graded.

Importantly, though, the probability of receiving a definitive grade is also significantly influenced by the location of the grade boundaries. In components where grade boundaries are close together (most likely because the assessment has not successfully spread out candidate marks), the marking consistency will have a more profound impact on the probability of being awarded the definitive grade. Thus, the wider the grade boundary locations, the greater the probability of candidates receiving the definitive grade. This is a very important point: the design of an assessment might be as important as marking consistency in securing the 'true' grade for candidates.

A summary statistic could be calculated by taking the mean of the probability that a candidate has been awarded the definitive grade. The red line is the weighted mean of the probabilities, where the weights are the number of candidates at each final mark. This approach would reflect the impact of quality of marking on the entry population and assumes that the seed items are representative of the cohort that took each component. In the absence of real mark distributions, the distribution of marks has crudely been assumed to be normally distributed around half the maximum mark of the component. In future work it is likely that mark distributions for each component will be used (either total mark distribution or mean and standard deviation) so the weighted probabilities will be more reflective of the actual cohort. It may also be desirable to present a statistic that is independent of mark distributions

and one which is only over the judgemental grade boundaries (as these are the grades most affected by quality of marking); this may be achieved by numerically integrating the region under the probability distributions from the lowest judgemental grade boundary to the highest grade boundary. This allows the area under the curve to be calculated and the entire process is explained in more detail by Press, Teukolsky, Vetterling, & Flannery, 2007. The resultant integral is subsequently divided by the number of marks separating the two boundaries. This is illustrated by the blue line in figure 9 and represents the probability that a candidate has been awarded the definitive grade over the judgemental grade boundaries only. This statistic generally represents the most conservative probability calculation because all marks outside of the judgemental grade boundaries are excluded; typically, the percentage of candidates who have been awarded the definitive grade is 100% within 2 or 3 marks outside this range. As a result, this statistic is always smaller than the probabilities weighted by mark distribution particularly for subjects where a large proportion of candidates are outside of the judgemental grade boundaries (for example AS physics where approximately 20% of candidates get an A).

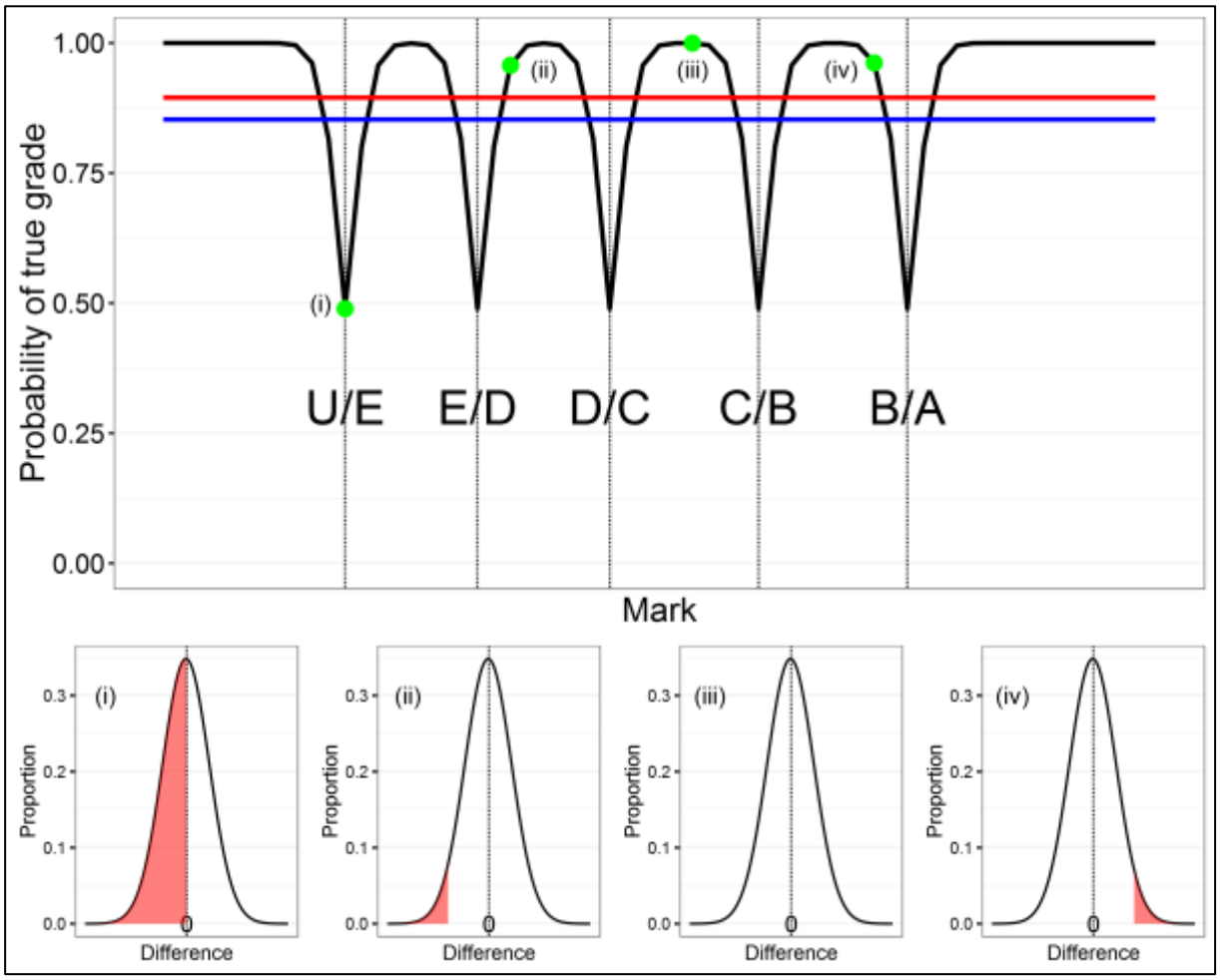

Figure 12. *The probability of being awarded the 'true'/definitive grade dependent on the final mark awarded to the candidate (solid black line) for a single AS physics component. It is observed that the probabilities dip at the grade boundaries. The probability at each mark is calculated from the proportion of candidates that are over- or under-graded and is illustrated at various points ((i), (ii), (iii) and (iv)). The distribution of differences at component level are shown in figures (i) to (iv) and the proportion of candidates not receiving the definitive grade is given by the shaded area. For example, on a grade boundary (figure (i)), any candidate awarded a mark more severe than the definitive mark will not receive the definitive grade. The probability weighted by the mark distribution is given by the solid red line; and the integrated probability given by the solid blue line.*

A comparison of the probability of being awarded the definitive grade for a selection of GCSE components within a single humanities subject is shown in figure 13. Relatively large variation is observed within this GCSE subject.

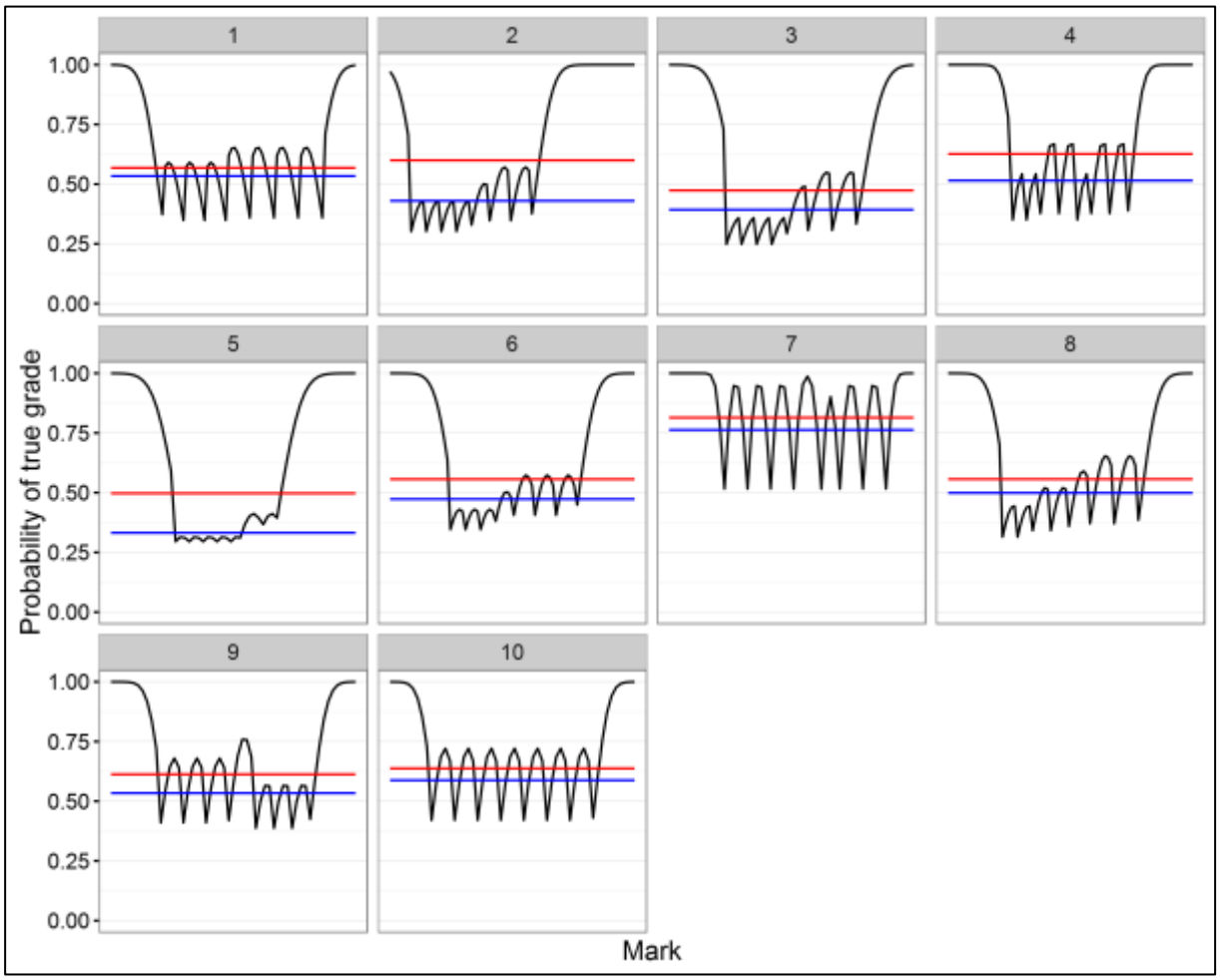

Figure 13. *The probability of being awarded the definitive grade dependent on the final mark awarded to the candidate for a selection of GCSE units within a single humanities subject.*

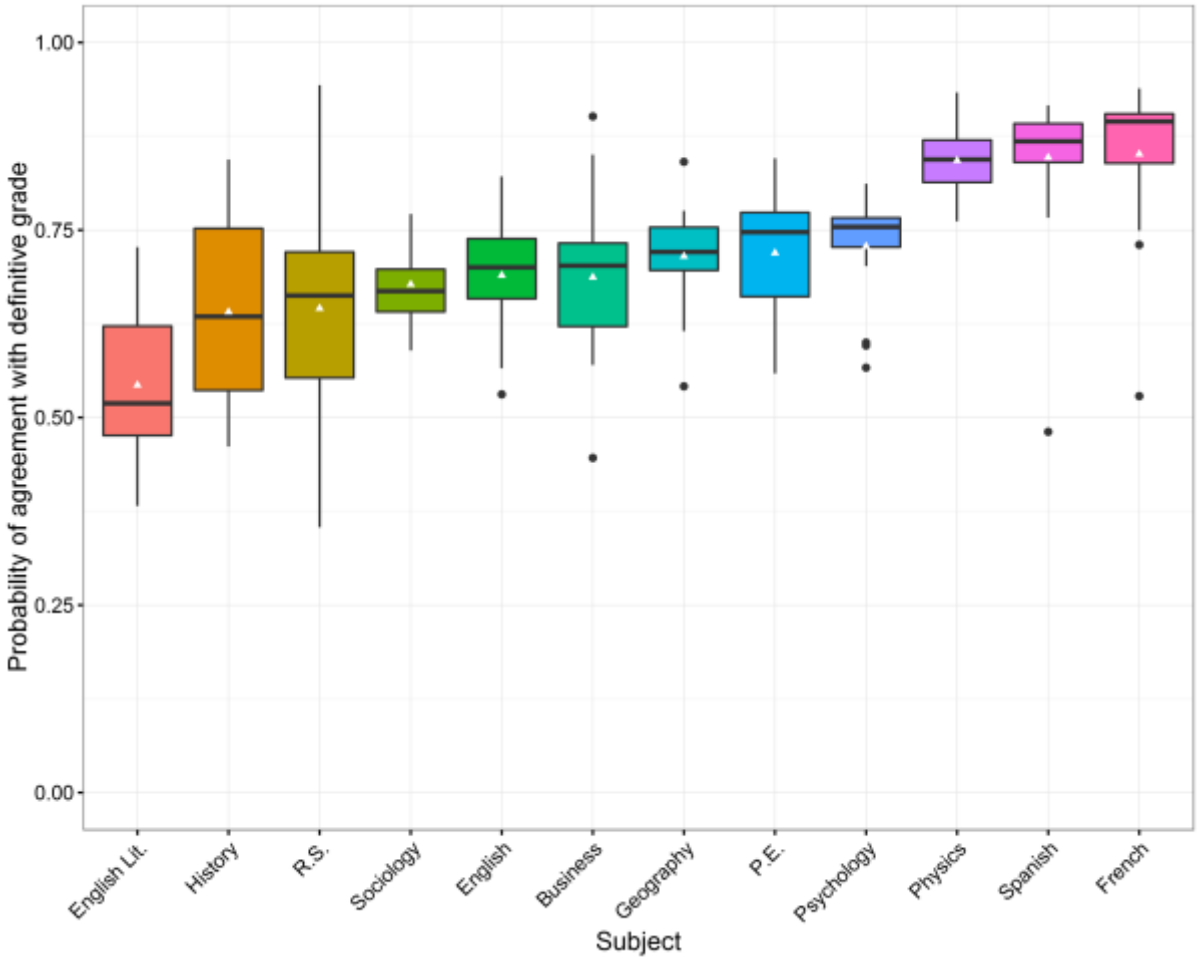

Figure 14. *Boxplot of the probability of a candidate being awarded the definitive grade. The mean probability for each subject is denoted by the white triangle.*

It is inevitable that there will be different levels of marking accuracy in different subject areas (figure 14). It is observed that the quality of marking for physics components is higher than that for the more 'subjective' English language or history components. Physics questions are generally low-mark tariff (≤ 6 marks) questions and typically there is an objectively correct answer to each question. For more subjective questions, there may be legitimate differences in applying the mark scheme between different examiners resulting in less agreement between examiners. Any future comparison between quality of marking metrics should therefore only be between closely related subjects; variation is to be expected between subjects but large variation within a subject is unlikely to be acceptable (provided mode of assessments, assessment objectives and content coverage are similar).

Multi-level modelling

The probability that a candidate is awarded the definitive grade can also be derived by using a multi-level model to fit quality of marking. The parameters from this model can then be used to simulate the mark-remark difference at component level for a set of randomly generated candidates. The algorithm for generating pseudo-candidates is discussed in more detail elsewhere (Schumann, E., 2009) and so will only be briefly covered here.

The multi-level model relates the mark-remark difference for each item to the final mark awarded to the candidate for the item and the maximum mark of the item within each component within each subject. These variables were chosen on the basis that they were both likely to influence the level of agreement between examiners. Indeed, evidence of the maximum mark affecting the agreement between the examiners can be seen in figure 6.

All components from a single subject from all exam boards are included in a single model. The model has been constructed with three levels. Marking events (*i*) are nested within questions (*j*) which are in turn nested within components (*k*). The model is given by the following equation:

$$diff_{ijk} = \beta_{0jk} + \beta_1 final\ mark_{ijk} + \beta_2 maximum\ mark_{ijk} + e_{ijk}. \quad (3)$$

For each question within a component, the final mark awarded is randomly generated for 5,000 candidates. For each question the distribution of marks is roughly uniform and the correlation between questions is approximately 0.4 (Schumann, E., 2009). Equation 3 is then used to simulate the randomly generated mark-remark difference for each candidate and this simulation is replicated 25 times, giving the equivalent of 125,000 candidates. Each candidate has a final mark for every question on the component and a corresponding mark-remark difference.

From this, the final mark awarded and mark-remark difference are calculated at component level for each candidate. For the final mark awarded to each candidate the mean difference and standard deviation are summarised from each of the 25 replications. Probabilities that a candidate has been awarded the definitive grade are determined by their positions relative to the nearest grade boundary. Finally, the output statistic is the weighted mean of the probability that a candidate has been awarded the definitive grade, where the weights are the number of pseudo-candidates at each mark.

A comparison of the probability a candidate is awarded the definitive grade calculated using the two different methodologies is shown in figure 15. Results are similar but it is worth reflecting on the differences between the two approaches. The random variable approach reflects only the existing data and no attempts have been made to extrapolate beyond. The expected differences and variances are calculated based on the seeds present. The probabilities derived from this data approximate quality of marking; they are representative of the chosen seed responses but not the whole population who entered the examination. This could potentially be overcome by use of the multi-level model as any missing values could be assigned, allowing for analysis of all potential outcomes. The probabilities calculated using the pseudo-candidates depend on the extent to which equation 3 correctly reflects the relationship between marker accuracy with the dependent variables; given that significantly more variation is observed when considering just the responses to seed items, it appears that the multi-level model should be refined further. This suggests that variables may be missing or the relationship between marker accuracy and the suggested variables may not be linear. Lastly, the characteristics of the pseudo-candidates depends heavily on the parameters fed into the algorithm (Schumann, E., 2009) but it is uncertain what these parameters should be and a comprehensive sensitivity analysis is required. If underlying mark distributions are requested in future, then the parameters of the algorithm can be amended accordingly so the mark distribution of pseudo-candidates closely agrees with the actual mark distributions.

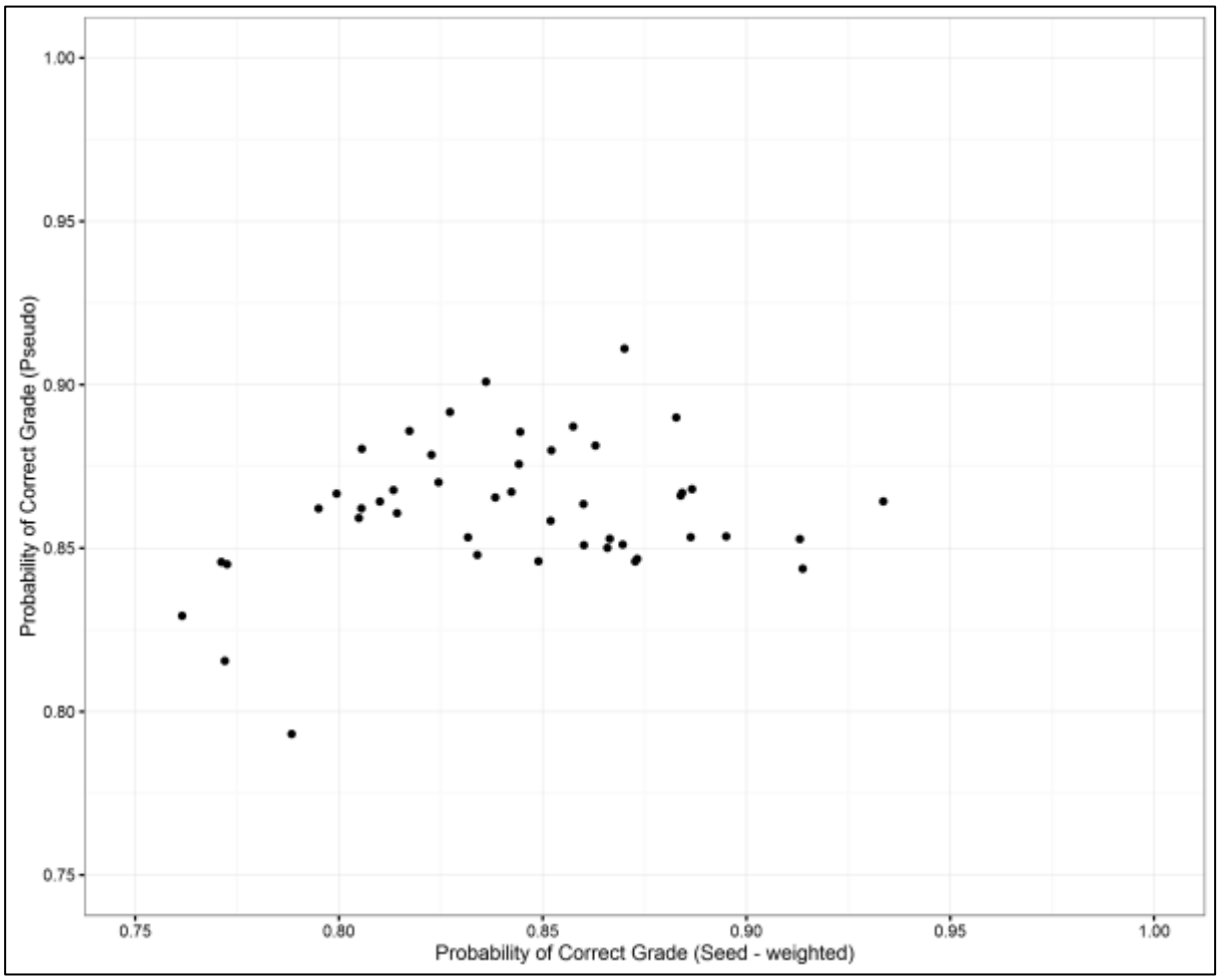

Figure 15. *Comparison between the two approaches of calculating the probability that a candidate obtains the definitive component grade.*

Given that marking accuracy, in its simplest form, is the difference between multiple marks, it is worth reflecting on whether quality of marking metrics should reference grade boundaries. After all, marking accuracy should be unrelated to the proximity of a given mark to a given grade boundary. However, this metric may be used to highlight components where a combination of poor marking and assessment design could lead to inaccurate grading; instances where grade boundaries were very narrow would exacerbate the effect of marking differences. If, historically, quality of marking was found to be stable within a component, then this metric could highlight the effect that changing the assessment would have on a particular component.

4.3 Specification level metrics

With the move from modular to linear assessment and a reduction in non-examined assessment as features of the new reformed GCSEs and A levels, it may be possible to derive specification level metrics by the aggregation of quality of marking for all

components within a specification. Such an approach has been applied to the AS and A level component level metrics for the 2015 physics data as an illustration of how these metrics may be used in future. Due to the lack of optionality in these physics components, the examples given are for one of the most straightforward scenarios.

Initially, specifications are grouped at qualification level, and, in-line with the changes to GCE, AS and A levels have been decoupled. The expected difference and standard deviations at specification level are estimated from all items within a specification in an approach identical to that at component level (equations 1 and 2). Any specifications where data from one or more externally assessed components are absent are excluded from this analysis.

The expected difference and standard deviation at specification level are illustrated in figure 16. Accuracy of marking is very similar between all specifications; the expected difference and standard deviation are typically found to be between ± 1% and 2% of the maximum mark of the specification respectively. These values have been used to generate the distribution of differences for pseudo-candidates illustrated in figure 17.

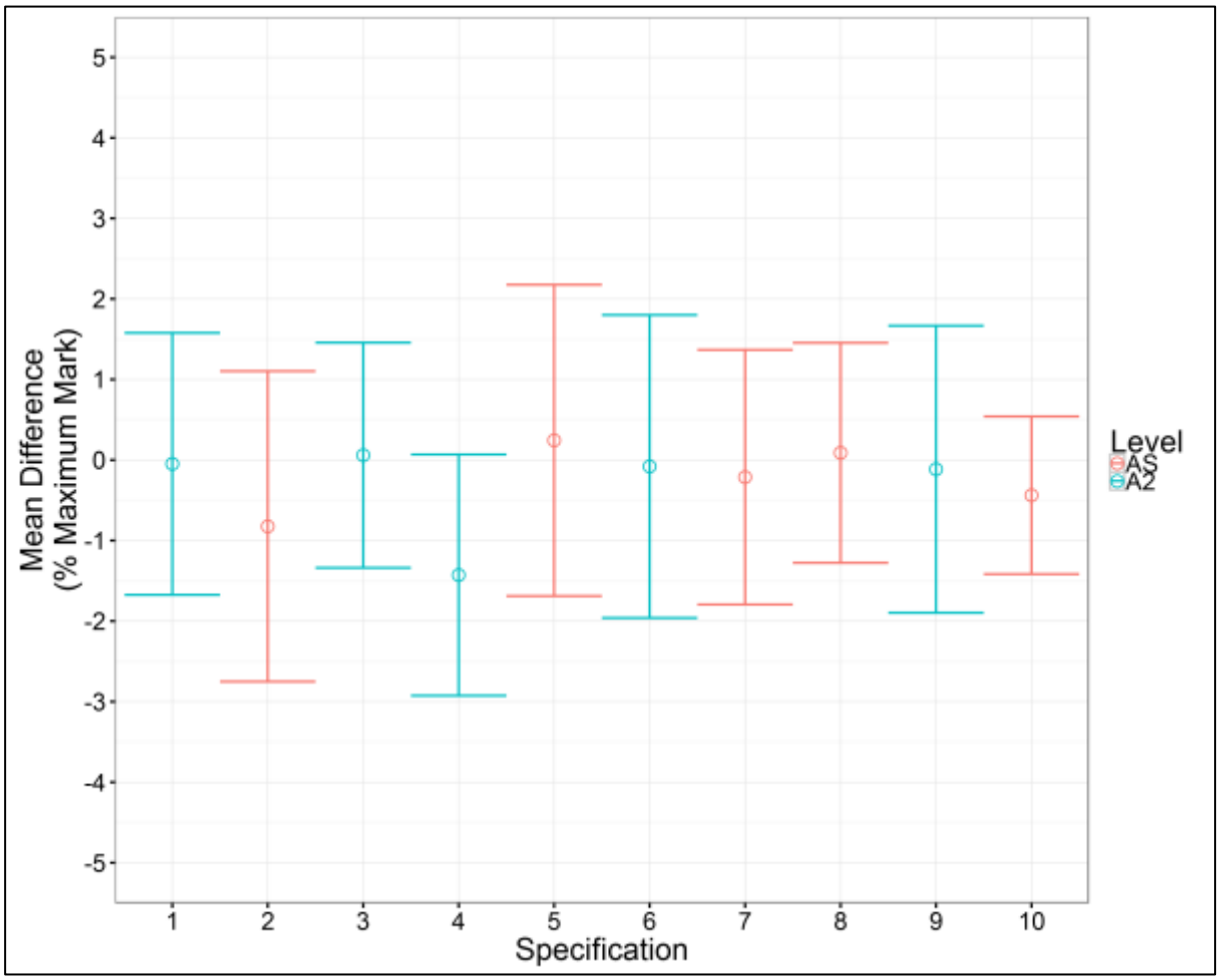

Figure 16. *The mean difference from the awarded mark expressed as a percentage of the maximum mark for each physics specification.*

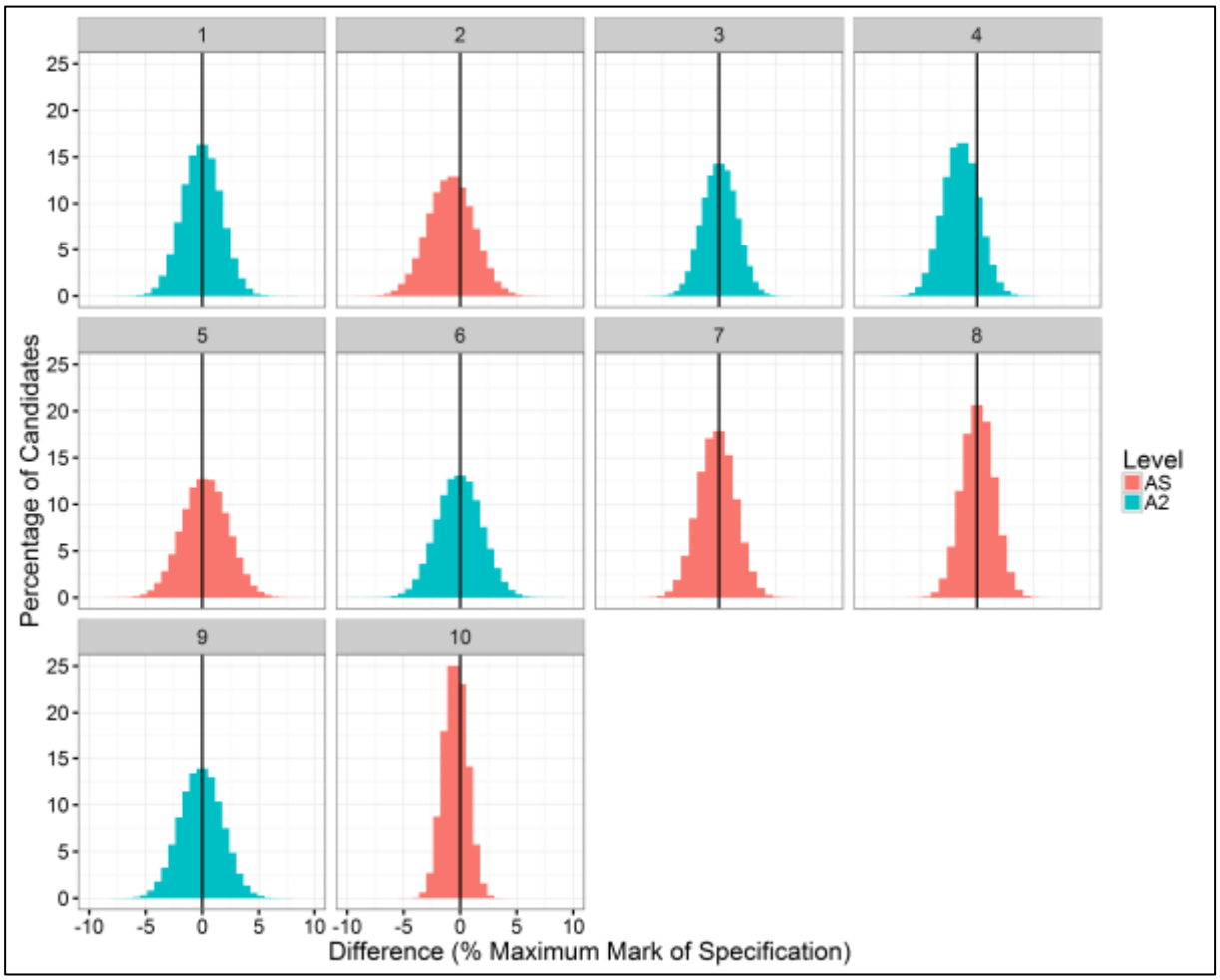

Figure 17. *The simulated difference from the awarded mark for each physics specification based on the randomly generated candidates.*

The aggregation of component grade boundaries into grade boundaries at specification level is significantly easier for linear assessments than modular ones. The specification level grade boundaries have been obtained by the simple addition of the component level grade boundaries. By using this approach, the probability that a candidate receives the definitive grade at qualification level may be calculated and is tabulated in table 3. When tabulated the data provides an opportunity to look at the difference that quality of marking and assessment design would have on the grade awarded at qualification level. In future, when data from the new GCSEs become readily available, the foundation and higher GCSE tiers could be combined.

By using the graphical presentation and tabulation of metrics it would also be possible to routinely summarise the output from all these metrics into a single page summary, allowing a quick comparison for all specifications from a suite of metrics. Combined, these metrics all highlight information that could be of interest.

Table 3. *Summary metrics for each physics specification.*

Specification	Expected Difference (% Max. Mark)	Standard Deviation (% Max. Mark)	Probability of definitive Grade (weighted)	Probability of definitive Grade (integration)
GCSE – F1	0.14	1.62	0.96	0.93
GCSE – F2	-0.63	3.23	0.89	0.83
GCSE – F3	-1.19	3.39	0.88	0.75
GCSE – F4	-0.90	2.60	0.90	0.83
GCSE – H1	-0.01	1.19	0.93	0.91
GCSE – H2	-0.47	1.95	0.86	0.81
GCSE – H3	-0.48	2.16	0.86	0.81
GCSE – H4	-0.48	1.38	0.90	0.88
AS – 1	0.09	1.36	0.91	0.83
AS – 2	-0.22	1.58	0.89	0.75
AS - 3	0.24	1.94	0.88	0.78
AS – 4	-0.83	1.94	0.86	0.69
AS – 5	-0.44	0.98	0.92	0.80
A2 – 1	0.06	1.39	0.89	0.82
A2 – 2	-0.05	1.62	0.88	0.77
A2 – 3	-0.09	1.88	0.88	0.75
A2 – 4	-0.13	1.78	0.89	0.76
A2 - 5	-1.43	1.49	0.85	0.59

5 Limitations

A series of item level and component level metrics have been derived from exam board data arising from on-line monitoring procedures. It has been necessary to make a series of assumptions in the derivation of these metrics. All the analysis in this report has assumed that the most appropriate basic measure of quality of marking is the difference between two independently awarded marks. In order to use the data from the exam boards, it has been necessary to assume

1. that the mark awarded to the seed item is the definitive (‘true’) mark. This is most likely the case for most seed items, but in instances where the most frequent mark awarded by examiners differs from the definitive mark there is a possibility that the definitive mark is wrong (Bramley & Dhawan, 2010). There are multiple approaches used for arriving at the definitive mark (Tisi, 2013) and, as there is no formal procedure for arriving at a single mark for a seed item, nor is there any formal recording of the process, it has been necessary to assume that no bias is introduced to the potential quality of marking metric by the way in which the final mark is derived.

2. that the two marks being compared are entirely independent. The assumption of independence is safer perhaps for sample double marking than for seeding items. In the latter case, for example, it is possible in some marking systems that in some cases those examiners involved in deriving the definitive mark for seeds are subsequently monitored using the same seeds. Additionally, it may be that examiners receive feedback (including the mark) on specific seeds and are able to retain and re-use this information if the same seed reappears subsequently. Where two marks are not independent, this would most likely provide an over-estimate of marking consistency for the purpose of these metrics[2].

There are other assumptions present in how metrics have been derived. It has been assumed that it is acceptable to collapse optional questions in the derivation of component and specification level metrics. This appears to be a reasonable assumption when comparing the distribution of differences between pseudo-candidates and actual candidate distribution. However, ideally perhaps, each optional route through a component or qualification should be treated as a separate entity (Stockford & He, 2014). Thus far specification level metrics have focussed on physics as there is very little optionality in these specifications. These metrics can be easily extended to specifications with complex optionality by use of pseudo-candidates. Following the question selection rules within each component in a specification, candidates can be generated taking each optional path within a component. Furthermore, once candidate data becomes available, probabilities can be assigned to each question based on the frequency at which it is chosen.

The metrics derived directly from the response to seed items reflect the data that is available. For example, if an item is worth 10 marks and the chosen seeds represent a range in marks from 3 to 7, then this is reflected in the expected difference for that particular item, which in turn is reflected in the metrics. Ideally, if on-screen marked data is to be used in the derivation of metrics, seeds should be selected across the entire mark range of the item, including zero and full-mark responses.

Due to the majority of onscreen marking being segmented, derivation of component level metrics is not trivial. In the pursuit of component or specification level metrics, ideally seeds would be at least script level, allowing for easy calculation of component level metrics. However, this would require operational, procedural and computational changes for the exam boards and would mean that the creation of

[2] It is also worth pointing out that any loss of independence of the two marks not only undermines the metrics but also undermines the true purpose of seeding which is to monitor live marking.

quality of marking metrics was prioritised over monitoring the quality of marking. Figure 6 also highlights the difficulty in using the data generated from on-line monitoring procedures in the creation of metrics. It would be very easy to artificially improve the quality of marking seen in these metrics by the removal of 'good' seeds with high failure rates, however this would come at the expense of a robust monitoring process. However, if seeds were chosen for a script/item that was difficult to mark (for example to check the examiners understanding of how to apply the mark scheme in such instances) and these seeds were over-represented, then these metrics would under-estimate the quality of marking for non-seed items (Bramley & Dhawan, 2010).

The probabilities calculated using the multi-level model require further refinements as this approach represents a first attempt at using a model to simulate the mark-remark difference. The fact that this approach shows less variation than the method using response to seed items suggests that the linear relationship between mark difference and the independent variable should be the subject of further scrutiny. A full sensitivity analysis is needed. It needs to consider the impact of varying the pseudo-candidate parameters and to explore the fit of the underlying multi-level model taking into account non-linear relationships, interactions and independent variables.

6 Conclusions and future work

A series of metrics are presented in this report as are the conditions necessary to derive them. After a review of the on-line monitoring process and exam board data, a series of item level statistics are derived which are used as the foundations of component level metrics. These metrics are presented in a manner to highlight differing aspects of quality of marking. After a series of simple assumptions, these metrics are then scaled up to specification level to give some indication of how they may be presented when the reforms brought in by the new GCSEs and A levels come into effect. Limitations with both the metrics and on-line marked data have been listed.

- The assumption that automarked items are marked perfectly accurately seems a reasonable assumption, however, this is likely to need further exploration.
- Optionality within components and specifications should be investigated further by use of pseudo-candidates.
- A full sensitivity analysis of the multi-level model is required.

- Some of the data supplied by the exam boards needs to checked. When mark-remark data is not available for one or more questions within a component, a procedure for dealing with the missing data is required.
- Given the complexity and sensitivity of the data it is essential that the metrics stand up to scrutiny and that there is a very clear understanding behind the meaning and application of any quality of marking metric. There are dangers that information from metrics (particularly when related to grade boundaries) could be used out of context.
- Most importantly, it is essential that metrics, or rather the *use* of these metrics, do not compromise the live on-line monitoring procedures. It would be beneficial to continue the programme of developing and refining metrics to test their robustness.

Further consideration of the practical uses of such metrics, including derivation of acceptable levels of marking consistency or how they might best be used to drive improvements in marking quality (without compromising the live on-line monitoring procedures) need exploration.

7 References

Bramley, T., & Dhawan, V. (2010). Reliability of qualifications report, *2907*(2907).

E. Ofqual (2014). Review of quality of marking in exams in A levels , GCSEs and other academic qualifications, (February).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge University Press.

Schumann, E. (2009). Generating Correlated Uniform Variates. Retrieved from http://comisef.wikidot.com/tutorial:correlateduniformvariates

Stockford, I., & He, Q. (2014). *Reporting of assessment functioning statistics for regulated qualifications - a paper for discussion*.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). A review of literature on marking reliability research (report for Ofqual), (June).

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346